# REPORT DOCUMENTATION PAGE

*Form Approved*
**OMB No. 0704-0188**



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| OCT 2007 | Journal Article Postprint | |

**4. TITLE AND SUBTITLE**

FREE ENERGY GAP AND STATISTICAL THERMODYNAMIC FIDELITY OF DNA CODES

**5a. CONTRACT NUMBER**
FA8750-07-C-0089

**5b. GRANT NUMBER**
N/A

**5c. PROGRAM ELEMENT NUMBER**
61102F

**6. AUTHOR(S)**

Morgan A. Bishop, Arkadii G. D'Yanchkov, Anthony J. Macula, Thomas E. Renz, and Vyacheslav V. Rykov

**5d. PROJECT NUMBER**
232T

**5e. TASK NUMBER**
DN

**5f. WORK UNIT NUMBER**
AC

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Macula, Inc.
36 Westview Crescent
Geneseo, NY 14454

**8. PERFORMING ORGANIZATION REPORT NUMBER**
N/A

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

AFOSR
875 North Randolph Street
Arlington, VA 22203

AFRL/RITC
525 Brooks Rd.
Rome NY 13441-4505

**10. SPONSOR/MONITOR'S ACRONYM(S)**
N/A

**11. SPONSORING/MONITORING AGENCY REPORT NUMBER**
AFRL-RI-RS-TP-2009-3

**12. DISTRIBUTION AVAILABILITY STATEMENT**

Approved for public release; distribution unlimited. PA# WPAFB 07-279 Date Cleared: Jun14, 2007.

**13. SUPPLEMENTARY NOTES**

 Paper published in the Journal of Computational Biology, V. 14, No. 8, October 2007, pp. 1088-1104. DOI: 10.1089/cmb. 2007.0083. 

**14. ABSTRACT**

DNA nanotechnology often requires collections of oligonucleotides called "DNA free energy gap codes" that do not produce erroneous cross hybridizations in a competitive multiplexing environment. This paper addresses the question of how to design these codes to accomplish a desired amount of work within an acceptable error rate. Using a statistical thermodynamic and probabilistic model of DNA code fidelity and mathematical random coding theory methods, theoretical lower bounds on the size of DNA codes are given. More importantly, DNA code design parameters (e.g., strand number, strand length and sequence composition) needed to achieve experimental goals are identified

**15. SUBJECT TERMS**

DNA Codes, cross hybridization, DNA design, random coding theory

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Thomas E. Renz |
| U | U | U | UU | 18 | 19b. TELEPHONE NUMBER (*Include area code*) N/A |

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 8, 2007

Pp. 1088–1104
DOI: 10.1089/cmb.2007.0083

# Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes

MORGAN A. BISHOP,[1] ARKADII G. D'YACHKOV,[2] ANTHONY J. MACULA,[3]
THOMAS E. RENZ,[4] and VYACHESLAV V. RYKOV[5]

## ABSTRACT

**DNA nanotechnology often requires collections of oligonucleotides called "DNA free energy gap codes" that do not produce erroneous crosshybridizations in a competitive muliplexing environment. This paper addresses the question of how to design these codes to accomplish a desired amount of work within an acceptable error rate. Using a statistical thermodynamic and probabilistic model of DNA code fidelity and mathematical random coding theory methods, theoretical lower bounds on the size of DNA codes are given. More importantly, DNA code design parameters (e.g., strand number, strand length and sequence composition) needed to achieve experimental goals are identified.**

**Key words:** biomolecular computing, crosshybridization, DNA barcodes, DNA codes, DNA computing, DNA words, hybridization, nearest neighbor, random coding methods, self assembly, stacked pairs, statistical thermodynamics, SynDCode, tag-antitag systems, universal arrays.

## 1. INTRODUCTION

DNA NANOTECHNOLOGY often requires collections of oligonucleotides that do not produce erroneous crosshybridizations. When these collections consist of complementary pairs of oligonucleotides, i.e., are closed under complementation, they are called *DNA tag-antitag systems* (Kaderali et al., 2003) and *DNA codes* (D'yachkov et al., 2003, 2005b, 2006). When the collections need not be closed under complementation they are called *DNA words* (Andronescu et al., 2003; Tulpan et al., 2005; Shortreed et al., 2005) and *DNA barcodes* (Eason et al., 2004). These collections of non-crosshybridizing collections have applications in SNP multiplexing (Cai et al., 2000; Kaderali et al., 2003; Fish et al., 2007), gene function identification (Eason et al., 2004), nanostructure self-assembly (Valignat et al., 2005), universal microarrays (Hardenbol et al., 2003), and biomolecular computing (Braich et al., 2002; Rose et al., 2004). Combinatorial, heuristic, and biological methods have been suggested as a means by which DNA codes can be found and programs exist that generate DNA codes (Andronescu et al., 2003; Bishop et al., 2006; Chen et al., 2006; Tulpan et al., 2005; D'yachkov et al., 2006; Penchovsky and Ackermann, 2003).

[1]JEANSEE, Geneseo, New York.
[2]Faculty of Mechanics and Mathematics, Department of Probability Theory, Moscow State University, Moscow, Russia.
[3]Biomathematics Group, SUNY Geneseo, Geneseo, New York.
[4]Air Force Research Laboratory, IFTC, Rome Research Site, Rome, New York.
[5]Department of Mathematics, University of Nebraska-Omaha, Omaha, Nebraska.

In several papers (Zhang et al., 2005; Tulpan et al., 2005; Dirks et al., 2004, 2007; Rose et al., 1999, 2004; Horne et al., 2006), statistical thermodynamics is applied to model competitive multiplexing hybridization. However, the methods there are primarily numerical in nature and do not provide detailed information about how to design collections of non-crosshybridizing strands to accomplish a desired amount of work within an acceptable error rate. This paper concerns exactly this question and presents a theoretical, not a heuristic, numerical or algorithmic, way to decide this question. Using a statistical thermodynamic and probabilistic model of DNA code fidelity coupled with mathematical random coding theory methods similar to those presented in D'yachkov et al. (2005b, 2003), theoretical lower bounds on the size of DNA codes are given. More importantly, DNA code design parameters, e.g., strand number, strand length and sequence composition, needed to achieve experimental self-assembly goals are identified.

Single strands of DNA are represented by *(A, C, G, T)* -quaternary sequences that are oriented, either $5' \to 3'$ or $3' \to 5'$. In this paper, *single stranded* DNA molecules without an indicated direction are assumed to be in the $5' \to 3'$ direction. The *reverse-complement* of a DNA strand is defined by first reversing the order of the letters and then substituting each letter with its complement, *A* for *T*, *C* for *G* and vice-versa. For example, the reverse complement of *AACGTG* is *CACGTT*. Henceforth, *complement* means reverse-complement unless otherwise stated. For strand $x$, let $\bar{x}$ denote its complement. A (perfect) *Watson-Crick duplex* is the joining of complement sequences in opposite orientations so that every base of one strand is paired with its complementary base on the other strand in the double helix structure, i.e., x and $\bar{x}$ are "perfectly compatible." However, when two, not necessarily complementary, oppositely directed DNA strands are "sufficiently compatible," they too are capable of coalescing into a double stranded DNA duplex. The process of forming DNA duplexes from single strands is referred to as *DNA hybridization. Crosshybridization* is when two oppositely directed and non-complementary DNA strands form a duplex. Crosshybridization doesn't always occur, but there is a potential for it to happen. In general, crosshybridization is undesirable as it usually leads to experimental error. To increase the accuracy and throughput of the applications listed above, there is a desire to have collections of DNA strands, as large and as mutually incompatible as possible, so that no crosshybridization can take place.

**Definition 1.** *Given two DNA strands x and y, we let $x : y$ denote the DNA duplex formed between x and y. It is implicitly assumed that x and y are oppositely oriented in $x : y$ with the first strand x* always *assumed to be in the $5' \to 3'$ direction and the second y always assumed to be in the $3' \to 5'$ direction. A crosshybridized (CH) duplex is an $x : y$, where $y \neq \bar{x}$.*

Even though it is possible for complementary sequences to form a non-perfectly aligned duplex, we will call any $x : \bar{x}$ duplex a Watson-Crick (WC) duplex. Two oppositely directed copies of a single strand $x$ can form $x : x$, which is a CH duplex if $x$ is not self-complementary, e.g., $x = ACGT = \bar{x}$. In the discussion below, self-complementary strands are largely forbidden.

## 2. STACKED PAIRS AND UNSTACKED 2-STRINGS IN SECONDARY STRUCTURES

Let $x = x_1, \ldots, x_n$ and $y = y_1, \ldots, y_n$ be DNA sequences. For a base $y_j$, let $\tilde{y}_j$ be its complement base. Then $\bar{y} = \tilde{y}_n, \ldots, \tilde{y}_1$.

**Definition 2.** *Suppose $1 \leqslant i_r, j_r \leqslant n$. A* secondary structure *of the DNA duplex $x : y$ is a sequence of pairs of complementary bases $\rho = (x_{i_r}, y_{n+1-j_r})$ where $x_{i_r} = \tilde{y}_{n+1-j_r}$ and $(x_{i_r})$ and $(y_{n+1-j_r})$ are increasing and decreasing subsequences of x and y respectively. Given a secondary structure $\rho = (x_{i_r}, y_{n+1-j_r})$ a* stacked pair *in a duplex is a pair of consecutively aligned complementary bases, $x_{i_r} = \tilde{y}_{n+1-j_r}$, $x_{i_r+1} = \tilde{y}_{n+1-j_r+1}$, in $\rho$ where $i_{r+1} = i_r + 1$ and $j_{r+1} = j_r + 1$. The notation $x_{i_r}x_{i_r+1}/\tilde{y}_{n+1-j_r}\tilde{y}_{n+1-j_r+1}$ is used to denote a stacked pair. An* unstacked 2-string *of x in a secondary structure $\rho = (x_{i_r}, y_{n+1-j_r})$ is a 2-string, $x_i x_{i+1}$, of x that is not part of any stacked pair in $\rho$.*

Clearly the $x : y$ can have many secondary structures and stacked pairs and unstacked 2-strings must be defined relative to a given secondary structure.

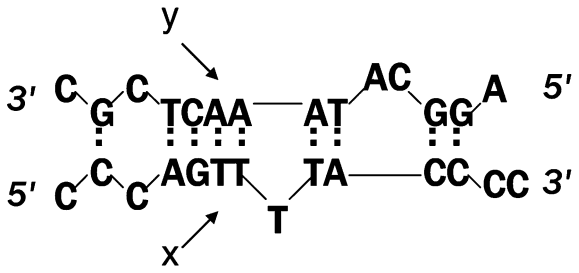


**FIG. 1.** An example of a secondary structure in a DNA duplex.

**Example 1.** *The secondary structure in Figure 1 has stacked pairs* F1

$$A_4G_5/T_{11}C_{10}, \quad G_5T_6/C_{10}A_9, \quad T_6T_7/A_9A_8, \quad T_9A_{10}/A_7T_6, \quad C_{11}C_{12}/G_3G_2$$

*where the subscripts indicate the position of the bases in the* $5' \to 3'$ *direction. Since*

$$A_4G_5, \quad G_5T_6, \quad T_6T_7, \quad T_9A_{10}, \quad C_{11}C_{12}$$

*are the* $5' \to 3'$ *bases in stacked pairs in the exhibited secondary structure, then the unstacked 2-strings in* $x$ *are*

$$C_1C_2, \quad C_2C_3, \quad C_3A_4, \quad T_7T_8, \quad T_8T_9, \quad A_{10}C_{11}, \quad C_{12}C_{13}, \quad C_{13}C_{14}.$$

**Definition 3.** *Given two DNA strands* $x$ *and* $y$*, let* $S(x, y)$ *and* $U(x, y)$ *respectively denote the maximum and minimum number of stacked pairs and unstacked 2-strings over all secondary structure between* $x$ *and* $y$ *and in* $x$*.*

A dynamic programming method to compute $S(x, y)$ is given in D'yachkov et al. (2005a, 2006). It is implemented in DNA code software *SynDCode* which is available at Bishop et al. (2006).

**Definition 4.** *Let* $L$ *be a collection of 2-strings of DNA bases closed under complementation, e.g.,* $L = \{AA, TT, AT, TA\}$ *or* $\{AT\}$*. A DNA sequence* $x = x_1, \ldots, x_n$ *is called an* L sequence of length n *if* $x_i x_{i+1} \notin L$ *for each* $1 \leqslant i \leqslant n-1$*. Let* $DNA(n, L)$ *denote the set* $L$ *sequences of length* $n$*. The cardinality of* $DNA(n, L)$ *is denoted by* $\lambda_{n,L}$ *or just* $\lambda_n$ *when the context is clear.*

Throughout this paper, $k, n, s, u$ and $N$ denote positive integers and, whenever $x \in DNA(n, L)$ is selected, it is assumed that each such $x$ is equally likely.

**Proposition 1.** *Let* $L = \{AA, TT, AT, TA\}$*, then* INS-A: BEGIN

$$\lambda_{n,L} = \frac{5-3\sqrt{5}}{10}\left(1-\sqrt{5}\right)^n + \frac{5+3\sqrt{5}}{10}\left(1+\sqrt{5}\right)^n \tag{2.1}$$

*and*

$$\lambda_{n,L} \leq 4\left(1+\sqrt{5}\right)^{n-1}. \tag{2.2}$$

**Proof.** A sequence in $DNA(n, L)$ can have at most $\lceil \frac{n}{2} \rceil$ of the letters $A$ or $T$ and no two of these can be consecutive. So given $x \in DNA(n, L)$, suppose the number of letters $A$ or $T$ it contains is $k$ where $0 \leqslant k \leqslant \lceil \frac{n}{2} \rceil$. There are $2^k$ different ways to arrange these. Then between any two of the letters $A$ or $T$ at least one $G$ or $C$ must be inserted. By a classical combinatorial "objects in boxes" type argument, there are

$$\binom{n-k-(k-1)+(k+1)-1}{n-k-(k-1)} 2^{n-k} = \binom{n-k+1}{k} 2^{n-k}$$

ways to insert the $n-k$ letters G or C. Thus

$$\lambda_{n,L} = 2^n \sum_{k=0}^{\lceil \frac{n}{2} \rceil} \binom{n-k+1}{k}.$$

It is known that

$$\sum_{k=0}^{\lceil \frac{n}{2} \rceil} \binom{n-k+1}{k} = F(n)$$

where $F(n)$ is recursively defined by $F(n) = F(n-1) + F(n-2)$ and $F(1) = 2$ and $F(2) = 3$. By the solving the Fibonacci recurrence relation with the given initial conditions $F(n) = \frac{5-3\sqrt{5}}{10}\left(\frac{1-\sqrt{5}}{2}\right)^n + \frac{5+3\sqrt{5}}{10}\left(\frac{1+\sqrt{5}}{2}\right)^n$. From this, (2.1) follows. Since

$$\begin{aligned}
\lambda_{n,L} &\leq \left(\frac{3\sqrt{5}-5}{10}\left(\sqrt{5}-1\right)^n + \frac{5+3\sqrt{5}}{10}\left(1+\sqrt{5}\right)^n\right) \\
&= \left(\frac{3\sqrt{5}-5}{10}\left(\frac{\sqrt{5}-1}{1+\sqrt{5}}\right)^n (1+\sqrt{5})^n + \frac{5+3\sqrt{5}}{10}\left(1+\sqrt{5}\right)^n\right) \\
&\leq \left(\frac{3\sqrt{5}-5}{10}\left(\frac{\sqrt{5}-1}{1+\sqrt{5}}\right)(1+\sqrt{5})^n + \frac{5+3\sqrt{5}}{10}\left(1+\sqrt{5}\right)^n\right) \\
&= 4\left(1+\sqrt{5}\right)^{n-1},
\end{aligned}$$

(2.2) is established. ■

**Definition 5.** *For $x, y, z \in DNA(n, L)$ with $x \neq y$, let*

$$B_{n,L,s}(x) \equiv \{y : S(x,y) \geqslant s\}.$$

$$A_{n,L,s} \equiv \{z : S(z,z) \geqslant s\}.$$

**Proposition 2.** *Let $L = \{AA, TT, AT, TA\}$ and $x, y, z \in DNA(n, L)$.*

*a.* $|B_{n,\varnothing,s}(x)| \leqslant \sum_{j=1}^{\min(s,n-s)} \binom{s-1}{j-1}\binom{n-s}{j}^2 4^{n-s-j}.$

*b.* $|B_{n,L,s}(x)| \leqslant \sum_{j=1}^{\min(s,n-s)} \binom{s-1}{j-1}\binom{n-s}{j} \min\left(4^{n-s-j}, 4^{j+1}\left(1+\sqrt{5}\right)^{n-s-2j-1}\right).$

*c.* $|A_{n,\varnothing,s}| \leqslant \sum_{\substack{j=1 \\ s+j \text{ even}}}^{\min(s,n-s)} \binom{\lfloor \frac{s}{2} \rfloor}{\lceil \frac{j}{2} \rceil - 1}\binom{n-s}{j} 4^{n-\frac{s+j}{2}}.$

*d.* $|A_{n,L,s}| \leqslant \sum_{\substack{j=1 \\ s+j \text{ even}}}^{\min(s,n-s)} \binom{\lfloor \frac{s}{2} \rfloor}{\lfloor \frac{j}{2} \rfloor - 1}\binom{n-s}{j} \min\left(4^{\frac{2n-s-j}{2}}, 4^{\lceil \frac{3j+2}{2} \rceil}\left(1+\sqrt{5}\right)^{\lfloor \frac{2n-s-4j-2}{2} \rfloor}\right).$

**Proof.** Proposition 2a appears as Proposition 12 in (D'yachkov et al., 2006) for $L = \emptyset$. A relatively straightforward modification of the proof there yields Proposition 2b here. Also see (D'yachkov et al., 2005b) where a proof of Proposition 2c appears.

To show Proposition 2b here, let $X, Y$ be binary sequences of length $n$ over $\{0, 1\}$ and $\{0, 2\}$ respectively. There are $\binom{s-1}{j-1}\binom{n-s}{j}^2$ pairs $X$ and $Y$ such that each have $s + j$ $0s$ with the $0s$ partitioned into $j$ substrings so that each substring has at least two $0s$ and these $0s$ are partitioned in exactly the same way in $X$ and $Y$. For example, $X = 000, 1, 00, 11, 0000$ and $Y = 2, 000, 00, 2, 0000, 2$ where the common partition of the $0s$ is $000, 00, 0000$.

Fix $x \in DNA(n, L)$ and suppose $y \in DNA(n, L)$ has $S(x, y) \geqslant s$. By the arguments in (D'yachkov et al., 2005a) and (D'yachkov et al., 2005b), for some $1 \leq j \leq \min(s, n - s)$ there is a common subsequence $(x_{i_k}) = (y_{j_k})$ of length $s + j$ of $x$ and $y$ respectively that can be obtained from a pair $X, Y$ above by taking $(x_{i_k}), (y_{j_k})$ to be the subsequences in $x$ and $y$ that correspond to the positions of the $0s$ is in $X, Y$ respectively.

Since $x$ is fixed, place $y$ in class $X, Y$ if the required common subsequence was obtained from this pair. The number of different $y$ in a given class is the number of $\{A, C, G, T\}$ sequences that can arise from a $y \in DNA(n, L)$ when it is restricted to the positions of the $2s$ in $Y$. The number of such subsequences that can arise is at most

$$\lambda_{b_1,L} \cdot \lambda_{b_2,L} \cdot \ldots \cdot \lambda_{b_{j+1},L}$$

where

$$\sum_{i=1}^{j+1} b_i = n - s - j \qquad \text{for } b_i \geq 0 \text{ and } \lambda_{0,L} \equiv 1.$$

From (2.2)

$$\lambda_{b_1,L} \cdot \lambda_{b_2,L} \cdot \ldots \cdot \lambda_{b_{j+1},L} \leq \min\left(4^{n-s-j}, 4^{j+1}\left(1 + \sqrt{5}\right)^{n-s-2j-1}\right)$$

and Proposition 2b then follows.

To show Proposition 2d here, let $Z$ be a binary sequence of length $n$ over $\{0, 1\}$. There are $\binom{\lfloor \frac{s}{2} \rfloor}{\lceil \frac{j}{2} \rceil - 1}\binom{n-s}{j}$ such $Z$ that have $s + j$ $0s$ with $s + j$ even and with the $0s$ symmetrically partitioned into $j$ substrings so that each substring has at least two $0s$ and the first $\frac{s+j}{2}$ $0s$ are arranged as the mirror image of the last $\frac{s+j}{2}$ $0s$. For example, e.g., $Z = 11, 000, 1, 000, 00, 111, 000, 000, 1$ where the symmetric partition of the $0s$ is $000, 000, 0|0, 000, 000$.

Let $z \in DNA(n, L)$ and suppose $S(z, z) \geqslant s$. By the arguments in (D'yachkov et al., 2005b), for some $1 \leq j \leq \min(s, n-s)$, there is a $Z$ as described above such that the positions of the $0s$ in $Z$ correspond to the positions of a self-complement subsequence of length $s + j$ in $z$. Since a self-complement subsequence of length $s + j$ is determined by its first $\frac{s+j}{2}$ entries, it follows that there are most

$$\min\left(4^{n-\frac{s+j}{2}}, 4^{j+\left\lceil \frac{j}{2} \right\rceil + 1}\left(1 + \sqrt{5}\right)^{\left(n-\frac{s+j}{2}\right)-\left(j+\left\lceil \frac{j}{2} \right\rceil + 1\right)}\right)$$

$$= \min\left(4^{\frac{2n-s-j}{2}}, 4^{\left\lceil \frac{3j+2}{2} \right\rceil}\left(1 + \sqrt{5}\right)^{\left\lfloor \frac{2n-s-4j-2}{2} \right\rfloor}\right)$$

$z$ in each class $Z$ because there are at $j + 1$ blocks of $1s$ and $\left\lceil \frac{j}{2} \right\rceil$ blocks of $0s$ for which there is choice to place substrings of $DNA(b_i, L)$ to construct sequences that capture all possible $z$ in class $Z$. ■

The following Corollary 1 is now trivial.

**Corollary 1.** *Let* $0 \leqslant s \leqslant n-1$. *Suppose* $L = \{AA, TT, AT, TA\}$. *Select* $x, y \in DNA(n, L)$ *with* $x \neq y$. *Then*

$$a.\ \Pr(S(x,y) \geqslant s) \leqslant \frac{1}{\lambda_n} \sum_{j=1}^{\min(s,n-s)} \binom{s-1}{j-1}\binom{n-s}{j}^2 \min\left(4^{n-s-j}, 4^{j+1}\left(1+\sqrt{5}\right)^{n-s-2j-1}\right).$$

$$b.\ \Pr(S(x,x) \geqslant s) \leqslant \frac{1}{\lambda_n} \sum_{\substack{j=1 \\ s+j\ even}}^{\min(s,n-s)} \binom{\lfloor \frac{s}{2} \rfloor}{\lceil \frac{j}{2} \rceil - 1}\binom{n-s}{j} \min\left(4^{\frac{2n-s-j}{2}}, 4^{\lceil \frac{3j+2}{2} \rceil}\left(1+\sqrt{5}\right)^{\lfloor \frac{2n-s-4j-2}{2} \rfloor}\right).$$

**Proof.** Part a follows from Proposition 2b. Part b follows from Proposition 2d. ■ INS-A: END

**Definition 6.** *Let* $U(x, y)$ *denote the minimum number of unstacked 2-strings in* $x$ *over all secondary structures between* $x$ *and* $y$.

If $x$ and $y$ have a maximum number of $s$ stacked pairs over all secondary structures, then there must be a minimum number of $n - s - 1$ unstacked 2-strings among the $x_1x_2, \ldots, x_{n-1}x_n$ that are not stacked. Thus, $U(x, y) \leqslant u$ if and only if $S(x, y) \geqslant n - u - 1$. So the next Corollary 2 follows from Corollary 1.

**Corollary 2.** *Let* $0 \leqslant u \leqslant n-1$. *Suppose* $L = \{AA, TT, AT, TA\}$. *Select* $x, y \in DNA(n, L)$ *with* $x \neq y$. INS-B: BEGIN
*Then*

$$a.\ \Pr(U(x,y) \leqslant u) \leq F_1(u,n) \tag{2.3}$$

*where*

$$F_1(u,n) \equiv \frac{1}{\lambda_n} \sum_{j=1}^{\min(u+1,n-u-1)} \binom{n-u-2}{j-1}\binom{u+1}{j}^2 \min\left(4^{u+1-j}, 4^{j+1}\left(1+\sqrt{5}\right)^{u-2j}\right).$$

$$b.\ \Pr(U(x,x) \leqslant u) \leq F_2(u,n) \tag{2.4}$$

*where*

$$F_2(u,n) \equiv \frac{1}{\lambda_n} \sum_{\substack{j=1 \\ n-u-1+j\ even}}^{\min(u+1,n-u-1)} \binom{\lfloor \frac{n-u-1}{2} \rfloor}{\lceil \frac{j}{2} \rceil - 1}\binom{u+1}{j} \min\left(4^{\frac{n+u+1-j}{2}}, 4^{\lceil \frac{3j+2}{2} \rceil}\left(1+\sqrt{5}\right)^{\lfloor \frac{n+u-4j-1}{2} \rfloor}\right). \quad \text{INS-B: END}$$

## 3. BOUNDS ON NEAREST NEIGHBOR THERMODYNAMICS

Stacked pairs play a special role in the *Nearest Neighbor (NN)* model of DNA duplex thermodynamics (SantaLucia, 1998; Zuker et al., 1999). Briefly, local thermodynamic functions $\Delta H$, $\Delta S$, which are essentially independent of temperature $T$, are experimentally found for stacked pairs and other secondary structure motifs and then are used in an additive fashion to predict global thermodynamic values for duplexes. Free energy, $\Delta G$, at a given temperature $T$ is derived from $\Delta H$, $\Delta S$ by

$$\Delta G = \Delta H - T\Delta S \tag{3.1}$$

One example of these local functions for stacked pairs is given in Table 1 (SantaLucia, 1998). A demonstration of how these local functions are used to make global predictions is given in Example 2. T1

TABLE 1. NEAREST NEIGHBOR THERMODYNAMIC VALUES FOR STACKED PAIRS

| *Stacked pair* $5' \to 3'/3' \to 5'$ | $\Delta H$ *kcal/mol* | $\Delta S$ *cal/°Kmol* | $\Delta G_{310°K}$ *kcal/mol* |
|---|---|---|---|
| AA/TT=TT/AA | −7.9 | −22.2 | −1.02 |
| AC/TG=GT/CA | −8.4 | −22.4 | −1.46 |
| AG/TC=CT/GA | −7.8 | −21.0 | −1.29 |
| AT/TA | −7.2 | −20.4 | −0.88 |
| CA/GT=TG/AC | −8.5 | −22.7 | −1.46 |
| CC/GG=GG/CC | −8.0 | −19.9 | −1.83 |
| CG/GC | −10.6 | −27.2 | −2.17 |
| GA/CT=TC/AG | −8.2 | −22.2 | −1.32 |
| GC/CG | −9.8 | −24.4 | −2.24 |
| TA/AT | −7.2 | −21.3 | −0.60 |

**Example 2.** *The $\Delta H$, $\Delta S$ of the WC duplex $x : \bar{x} = AGTCA{:}TCAGT$ predicted by the NN model is computed by essentially summing associated Table 1 values for the duplex's stacked pairs,*

$$AG/TC, \quad GT/CA, \quad TC/AG, \quad CA/GT,$$

*then adding a constant initiation penalty* IP. *Thus for AGTCA-TCAGA,*

$$\Delta H(duplex) = -7.8 - 8.4 - 8.2 - 8.5 + IP_1 = -32.9 + IP_1 \text{ kcal/mol}$$

$$\Delta S(duplex) = -0.0210 - 0.0224 - 0.0222 - 0.0227 + IP_2 = -0.0883 + IP_2 \text{ kcal/°Kmol.}$$

*From these computed values, the free energy for the WC duplex is given by (3.1) and thus for the duplex at* $310°K$,

$$\Delta G(duplex) = -5.51 + IP_1 + IP_2 \text{ kcal/mol.}$$

*Another way to accomplish the same task is to first compute the $\Delta G$ for each stacked pair at a given temperature, sum these values and add $IP = IP_1 + IP_2$. The $\Delta G$ for stacked pairs at* $310°K$ *is given in the last column of Table 1.*

Predictions about the thermodynamic stability of CH duplexes with a given secondary structure can also be made from the stacked pairs that it contains. In D'yachkov et al. (2005a, 2006), it is argued that the $\Delta G$ for a CH duplex is bounded below the sum of all the free energies of the stacked pairs that it contains plus *IP*. *Note, the more negative $\Delta G$ means more stable*. The claim is strongly supported by comparing this thermodynamically weighted stacked pair sum computed by *SynDCode* (Bishop et al., 2006) to computations made by NN minimum free energy software *Pairfold* (Andronescu et al., 2003) and *RNAstructure* (Mathews et al., 2006).

**Example 3.** *The secondary structure in Figure 1 has stacked pairs*

$$A_4G_5/T_{11}C_{10}, \quad G_5T_6/C_{10}A_9, \quad T_6T_7/A_9A_8, \quad T_9A_{10}/A_7T_6, \quad C_{11}C_{12}/G_3G_2.$$

*Hence at 310°K and using values from Table 1, the $\Delta G$ of the duplex with the indicated secondary structure is bounded below by*

$$-1.29 - 1.46 - 1.02 - 1.60 - 1.83 = -7.20 + IP \text{ kcal/mol.}$$

*Thus using the standard $IP = 1.96$, SynDCode (Bishop et al., 2006) gives $\Delta G(duplex) = -6.24$ kcal/mol. Pairfold (Andronescu et al., 2003; Mathews et al., 2006) respectively give $-2.48$ and $-2.70$ kcal/mol.*

## 4. RELATIVE STABILITY AND PROBABILITY VIA STATISTICAL THERMODYNAMICS

One question is: How does one measure relative stability of DNA duplexes? For example, given *only* $N$ possible secondary $\rho_1, \ldots, \rho_N$ structures for a duplex with free energy $\Delta G_1, \ldots, \Delta G_N$ respectively where each $\Delta G_i \leqslant 0$, how likely is the duplex to have $\rho_i$? Thinking of the secondary structures as *states*, a statistical thermodynamic partition function argument can be applied. Let $\omega_i$ be some real-valued function of $\Delta G_i$. Then the partition function $Q$ is given by:

$$Q = \sum_{i=1}^{N} \omega_i. \tag{4.1}$$

Using $Q$, the probability that the duplex has secondary structure $\rho_i$ is given by

$$\Pr(\rho_i) = \frac{\omega_i}{Q}. \tag{4.2}$$

Typically, $\omega_i = \exp(|\Delta G_i|/RT)$ where $R = 0.0019872$ kcal/°Kmol is the gas constant and $T$ is temperature in degrees Kelvin. Following this thread:

$$\Pr(\rho_1) = \frac{\exp(|\Delta G_1|/RT)}{\sum_{i=1}^{N} \exp(|\Delta G_i|/RT)} = \frac{1}{1 + \sum_{i \neq 1} \exp(-(|\Delta G_1| - |\Delta G_i|)/RT)}. \tag{4.3}$$

If for all $i \neq 1$, $|\Delta G_1| - |\Delta G_i| \geqslant g \geqslant 0$. Then

$$\Pr(\rho_1) \geqslant \frac{1}{1 + (N-1)\exp(-g/RT)}. \tag{4.4}$$

This leads to the use of the *minimum free energy gap* as a measure of performance. This is explored in the next section (Tulpan et al., 2005; Penchovsky and Ackermann, 2003; Shortreed et al., 2005).

## 5. STATISTICAL THERMODYNAMIC SIGNIFICANCE OF THE FREE ENERGY GAP

Since each possible stacked pair can be identified with its 2-string in the $5' \to 3'$ strand (usually $x$) in the duplex, the absolute values of thermodynamic parameters in Table 1 are a function of 2-strings of bases. These positive functions are given in Table 2 where $\Delta H$, $\Delta S$ are renamed as $f_H$ and $f_S$ respectively. T2

TABLE 2. POSITIVE THERMODYNAMIC FUNCTIONS ON TWO STRINGS

| *Two-string* $5' \to 3'$ | $f_H$ | $f_S$ | $f_{G,310}$ |
|---|---|---|---|
| AA=TT | 7.9 | 22.2 | 1.02 |
| AC=GT | 8.4 | 22.4 | 1.46 |
| AG=CT | 7.8 | 21.0 | 1.29 |
| AT | 7.2 | 20.4 | 0.88 |
| CA=TG | 8.5 | 22.7 | 1.46 |
| CC=GG | 8.0 | 19.9 | 1.83 |
| CG | 10.6 | 27.2 | 2.17 |
| GA=TC | 8.2 | 22.2 | 1.32 |
| GC | 9.8 | 24.4 | 2.24 |
| TA | 7.2 | 21.3 | 0.60 |

The associated absolute value of the free energy for stacked pairs, $f_{G,T}$, at a given temperature $T$ can be computed by using the following version of the classical equation (3.1):

$$f_{G,T} = T \cdot f_S - f_H. \tag{5.1}$$

For example, $f_{G,310}$ is given in the last column of Table 2. Note that each of the functions $f_H$, $f_S$ and $f_{G,T}$ have the property that $f(x_i x_{i+1}) = f(\tilde{x}_{i+1}\tilde{x}_i)$, i.e., they are invariant under complementation.

*Henceforth $f_H$, $f_S$ will be arbitrary positive functions on 2-strings invariant under complementation and $f_{G,T}$ will be assume to have been derived from such via (5.1).*

**Definition 7.** *Given distinct and non-self-complementary strands $x$, $y$ in DNA$(n, L)$, consider the WC duplex $x : \bar{x}$ and the CH duplexes, $x : x$, $x : y$. Let $\|x, \bar{x}\|$ denote the absolute value of the $\Delta G$ of the WC duplex $x : \bar{x}$ in perfect alignment and let $\|x, x\|$ and $\|x, y\|$ denote the absolute values of the $\Delta G$ of most stable secondary structures for the $x : x$, $x : y$ CH duplexes The quantity*

$$\|x, \bar{x}\| - \|x, y\|$$

*is plainly referred to as the* asymmetric free energy gap.

Given distinct and non-self-complementary strands $x_1, x_2, \ldots, x_N$ in $DNA(n, L)$, think of strand $x_1$ having $N + 1$ possible states. It can either form a WC duplex with its complement $\overline{x}_1$ or it can form a CH duplex with one of the other $N$ strands $x_1, x_2, \ldots, x_N$ (itself included as there may be multiple copies of each strand.) Let $x = x_1$. Then following the partition function argument given in (4.3) and (4.4), the probability that $x$ forms a WC duplex is

$$\Pr(x : \bar{x}) = \frac{\exp(\|x : \bar{x}\|/RT)}{\exp(\|x : \bar{x}\|/RT) + \exp(\|x : x\|/RT) + \sum_{i=2}^{N} \exp(\|x : x_i\|/RT)}. \tag{5.2}$$

Now suppose that for $i \neq 1$, $\|x, \bar{x}\| - \|x, x\|$ and $\|x, \bar{x}\| - \|x, x_i\|$ are all at least $g \geqslant 0$. Then by (4.4) and (5.2),

$$\Pr(x : \bar{x}) \geqslant \frac{1}{1 + (N + 1)\exp(-g/RT)}. \tag{5.3}$$

Suppose $C = \{\{x_i, \bar{x}_i\}\}_{i=1}^{N}$ is a collection of $N$ complementary pairs of strands in $DNA(n, L)$ of multiplicity 2, such that the $2N$ strand types are distinct and thus no strand type is self-complementary. Then there are a total of $4N$ strands in solution. Suppose further that $\|x_i, \bar{x}_i\| - \|x_i, y\| \geqslant g \geqslant 0$ for all $y = x_j$ or $y = \overline{x}_j$ where $j \neq i$. Then for any strand $x_i$:

$$\Pr(x_i : \bar{x}_i) = \Pr(\bar{x}_i : x_i) \geqslant \frac{1}{1 + (2N - 1)\exp(-g/RT)}. \tag{5.4}$$

Under the reasonable assumption that the formation of the $x_i : \bar{x}_i$ duplex is independent (or doesn't reduce the probability) of the formation of the $x_j : \bar{x}_j$ duplex, then:

Given the entire collection $C$ *of* $4N$ strands, *the probability that* $2N$ *WC duplexes form so that there are no CH duplexes is at least*

$$\left(\frac{1}{1 + (2N - 1)\exp(-g/RT)}\right)^{2N}. \tag{5.5}$$

If (5.5) is to be at least $\alpha$ where $0 < \alpha < 1$, then solving for $g$ in terms of $\alpha$ and $N$ yields

$$g = -RT \ln\left(\frac{\alpha^{-1/2N} - 1}{2N - 1}\right). \tag{5.6}$$

**Definition 8.** *Given $f_{G,T}$ and any complementary pair $x : \bar{x}$, define*

$$\|x, \bar{x}\|_{f_{G,T}} \equiv \sum_{i=1}^{n-1} f_{G,T}(x_i x_{i+1}). \tag{5.7}$$

*Suppose $x$ and $y$ are non-complementary strands. Given a secondary structure $\rho = (x_{j_r}, y_{n+1-j_r})$ between $x$ and $y$, let $A(\rho)$ be the stacked pairs in $\rho$. Define*

$$\|x, y\|_{f_{G,T,\rho}} \equiv \sum_{x_{j_r} x_{j_r+1} \in A(\rho)} f_{G,T}(x_{j_r} x_{j_r+1}). \tag{5.8}$$

$$\|x, y\|_{f_{G,T}} \equiv \max_{\rho}(\|x, y\|_{f_{G,T},\rho}). \tag{5.9}$$

$$\gamma_{f_{G,T}}(x, y) \equiv \|x, \bar{x}\|_{f_{G,T}} - \|x, y\|_{f_{G,T}} \tag{5.10}$$

*$\gamma_{f_{G,T}}(x, y)$ is called the* asymmetric stacked pair free energy gap. *$\gamma$ is written for $\gamma_{f_{G,T}}$ when the context is clear.*

**Proposition 3.** *Let $x, y \in DNA(n, L)$. Suppose $0 \leqslant c \leqslant f_{G,T}$ for all $x_i x_{i+1} \notin L$. Then $\gamma_{f_{G,T}}(x, y) \geq U(x, y) \cdot c$*

**Proof.** Let $\rho = (x_{j_r}, y_{n+1-j_r})$ be a a secondary structure between $x$ and $y$, let $B_x(\rho)$ be the unstacked pairs in $x$ relative to $\rho$. From (5.7)–(5.11), it follows that

$$\gamma_{f_{G,T}}(x, y) = \sum_{x_{k_r} x_{k_r+1} \in B_x(\rho)} f_{G,T}(x_{k_r} x_{k_r+1})$$

for some $\rho$. Since for every $\rho$, $|B_x(\rho)| \geq U(x, y)$, the result follows. ■

Note $\gamma(x, \bar{x}) = 0$ and

$$\|x, y\|_{f_{G,T}} = \|y, x\|_{f_{G,T}}$$

$$\|x, y\|_{f_{G,T}} = \|\bar{x}, \bar{y}\|_{f_{G,T}}$$

so

$$\gamma(x, y) = \gamma(\bar{x}, \bar{y}).$$

However, since

$$\|x, \bar{x}\|_{f_{G,T}} \neq \|y, \bar{y}\|_{f_{G,T}}$$

then

$$\gamma(x, y) \neq \gamma(y, x)$$

and this is why the term "asymmetric" is used.

In D'yachkov et al. (2005a, 2006), it is discussed how $\|x, y\|_{f_{G,T}} - IP$ is an upper bound on $\|x, y\|$ when $f_H, f_S$ are as given in Table 2. The NN model gives that $\|x, \bar{x}\| = \|x, \bar{x}\|_{f_{G,T}} - IP$, so it follows that

$$\gamma(x, y) \leqslant \|x, \bar{x}\| - \|x, y\|. \tag{5.11}$$

Thus the asymmetric *stacked pair* free energy gap is a lower bound for the asymmetric free energy gap.

Definition 9 is central to this paper.

**Definition 9.** *Let $f_{G,T}$ be as given in equation (5.1) for some $f_H$, $f_S$. A $DNA_{f_{G,T}}(n, L, g)$ code $C$ is a collection of complementary pairs in $DNA(n, L)$ such that no strand is self-complementary and for any two non-complementary strands $x$, $y$ in $C$:*

$$\gamma_{f_{G,T}}(x, y) \geqslant g.$$

*$DNA_{f_{G,T}}(n, L, g)$ codes are also called* free energy gap DNA codes.

The DNA code software *SynDCode* (Bishop et al., 2006) generates $DNA_{f_{G,T}}(n, L, g)$ codes $C$ with many additional and optional user added sequence constraints (see Example 7 and Conclusion). It should also be noted that for $x, y \in C$ that since $\gamma(x, y) \geq g$ and $\gamma(y, x) \geq g$ that

$$\min(\|x, \bar{x}\|, \|y, \overline{y}\| - \|x, y\|) \geq g. \tag{5.12}$$

Thus $DNA_{f_{G,T}}(n, L, g)$ codes are *nearly* of the type discussed in Tulpan et al. (2005), Penchovsky and Ackermann (2003), and Shortreed et al. (2005), and in which (5.12) is *nearly* one of the of main constraints (see Conclusion).

## 6. RANDOM CODING BOUND FOR HIGH FIDELITY $DNA_{f_{G,T}}(n, L, g)$ CODES

Equations (5.6) and (5.11) provide a relationship between the free energy gap and the probability of correct self-assembly.

**Definition 10.** *Suppose only the strands of a $DNA_{f_{G,T}}(n, L, g)$ code $C$ are present in solution in equal concentrations. For $0 < \alpha < 1$, then $C$ self-assembles with fidelity $\alpha$ if $\alpha$ is the probability that every strand in $C$ forms a WC duplex.*

If the model of having exactly two copies of the strands of a $DNA_{f_{G,T}}(n, L, g)$ code $C$ is assumed to be reasonable for strands in equal concentrations, then from equation (5.11), a $DNA_{f_{G,T}}(n, L, g)$ code $C$ with $N$ pairs has fidelity $\alpha$ if, for any two non-complementary strands $x$, $y$ in $C$, the asymmetric free energy gap $\gamma(x, y)$ satisfies:

$$\gamma(x, y) \geqslant -RT \ln\left(\frac{\alpha^{-1/2N} - 1}{2N - 1}\right) = g. \tag{6.1}$$

Below a random coding theory method is used to get a lower bound on the number of complementary pairs $N$ of a $DNA_{f_{G,T}}(n, L, g)$ code.

**Example 4.** *If $C$ is a $DNA_{f_{G,T}}(n, L, g)$ with $N$ pairs of multiplicity 2 where $g$ is given by (6.1) when $\alpha = 1 - \frac{1}{2N}$, then out of the desired $2N$ WC duplexes, an expected number of one WC duplex does not form. In Figure 2, a sufficient stacked pair free energy gap $g(\alpha, N)$ with $\alpha = 1 - \frac{1}{2N}$ is plotted against $log_{10}(N)$.* F2

**Definition 11.** *Let $\mathcal{U} \equiv \{\{x, \bar{x}\} : \{x, \bar{x}\} \in DNA(n, L)\}$. An $E \subseteq \mathcal{U}$ is called a random $DNA(n, L)$ k-set of pairs if $|E| = k$ and the uniform distribution is on the k-sets of $\mathcal{U}$.*

For the remainder of this paper $E$ is assumed to be a random $DNA(n, L)$ k-set of pairs.

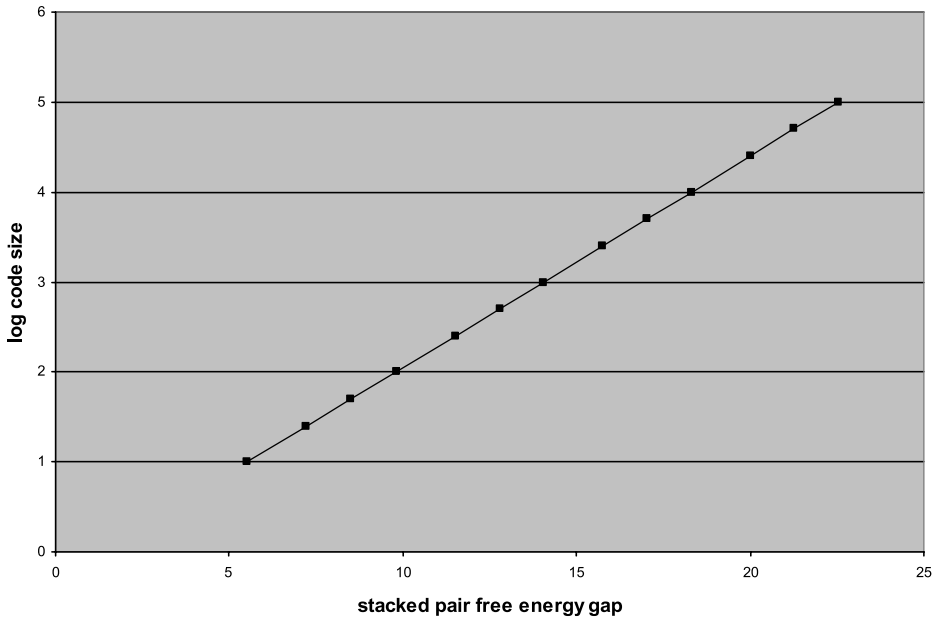


**FIG. 2.** Stacked pair free energy gap versus code size.

**Definition 12.** *For a real number $g > 0$, we say that a complementary pair $\{x_i, \bar{x}_i\}$ is $\gamma_{f_{G,T}}$ g-bad in $E$ if for any other $\{x_j, \bar{x}_j\}$ in $E$ either:*

$$\gamma_{f_{G,T}}(x_i, x_i) = \gamma_{f_{G,T}}(\bar{x}_i, \bar{x}_i) < g, \tag{6.2}$$

$$\gamma_{f_{G,T}}(x_i, x_j) = \gamma_{f_{G,T}}(\bar{x}_i, \bar{x}_j) < g, \tag{6.3}$$

$$\gamma_{f_{G,T}}(x_i, \bar{x}_j) = \gamma_{f_{G,T}}(\bar{x}_i, x_j) < g. \tag{6.4}$$

*A complementary pair $\{x_i, \bar{x}_i\}$ is called $\gamma_{f_{G,T}}$ g-good in $E$ if it* is not *$\gamma_{f_{G,T}}$ g-bad.*

The following is obvious:

**Proposition 4.** *The collection of $\gamma_{f_{G,T}}$ g-good pairs in $E$ is an $DNA_{f_{G,T}}(n, L, g)$ code.*

**Lemma 1.** *Let $x, y \in DNA(n, L)$ be randomly selected without replacement. Then there exists a $DNA_{f_{G,T}}(n, L, g)$ code with $N$ complementary pairs of strands from $DNA(n, L)$ if* INS-C: BEGIN

$$\Pr(\gamma(x, x) < g) + (2N - 1)\Pr(\gamma(x, y) < g) \leqslant \tfrac{1}{2}. \tag{6.5}$$

**Proof.** Let $\beta_g(x, y)$ be the probability that either $\gamma(x, y) < g$ *or* $\gamma(x, \bar{y}) < g$. Since $\Pr(\gamma(x, y) < g) = \Pr(\gamma(x, \bar{y}) < g)$ , then

$$\beta_g(x, y) \leqslant 2\Pr(\gamma(x, y) < g). \tag{6.6}$$

Let

$$\beta_g(x, x) \equiv \Pr(\gamma(x, x) < g). \tag{6.7}$$

Let $|E| = 2N$. From the additive bound of the probability of the union of events, it follows that the probability that $\{x_i, \bar{x}_i\}$ is $\gamma$ g-bad in $E$ is at most

$$\beta_g(x, x) + (2N - 1)\beta_g(x, y). \tag{6.8}$$

Thus the probability that $\{x_i, \bar{x}_i\}$ is good in $E$ is

$$1 - \left(\beta_g(x,x) + (2N-1)\beta_g(x,y)\right). \tag{6.9}$$

The main point of all of this is that the *expected number of good pairs in E is*

$$2N\left(1 - \left(\beta_g(x,x) + (2N-1)\beta_g(x,y)\right)\right). \tag{6.10}$$

and (6.10) will be at least $N$ when

$$\beta_g(x,x) + (2N-1)\beta_g(x,y) \leqslant \frac{1}{2}. \tag{6.11}$$

Thus given (6.11), there *must exist* an $E$ that contains $N$ $\gamma_{f_{G,T}}$ g-good pairs in $E$. Hence the result follows from Proposition 4. ■

**Proposition 5.** *Given $f_{G,T}$, suppose $0 \leqslant c \leqslant f_{G,T}$ for all $x_i x_{i+1} \notin L$. Let $g > 0$ and suppose that $\frac{g}{c}$ is not an integer. Then for distinct $x, y \in DNA(n, L)$*

$$a.\ \Pr(\gamma_{f_{G,T}}(x,y) < g) \leqslant F_1(\left\lfloor \frac{g}{c} \right\rfloor, n). \tag{6.12}$$

$$b.\ \Pr(\gamma_{f_{G,T}}(x,x) < g) \leqslant F_2(\left\lfloor \frac{g}{c} \right\rfloor, n). \tag{6.13}$$

**Proof.** Applying Proposition 3, it follows that

$$\Pr(\gamma_{f_{G,T}}(x,y) < g) \leqslant \Pr(U(x,y) \leqslant \left\lfloor \frac{g}{c} \right\rfloor). \tag{6.14}$$

The result follows from Corollary 2. ■

**Theorem 1.** *Let $L = \{AA, TT, AT, TA\}$. Given $f_{G,T}$, suppose $0 \leqslant c \leqslant f_{G,T}$ for all $x_i x_{i+1} \notin L$. Let $g > 0$ and suppose that $\frac{g}{c}$ is not an integer. Let $\delta_1(N,n,g,c)$ and $\delta_2(N,n,g,c)$ be as given in Proposition 5. Define $Bad_L(N,n,g,c)$ to be:*

$$Bad_L(N,n,g,c) \equiv F_2(\left\lfloor \frac{g}{c} \right\rfloor, n) + (2N-1)F_1(\left\lfloor \frac{g}{c} \right\rfloor, n). \tag{6.15}$$

*If $Bad_L(N,n,g,c) \leqslant \frac{1}{2}$, then there exists a $DNA_{f_{G,T}}(n,L,g)$ code with $N$ complementary pairs.*

**Proof.** Apply Lemma 1 and Propositions 4 and 5. ■ INS-C: END

## 7. RESULTS

In the Examples 5–7 below, let $L_1 = \{AA, TT, AT, TA\}$ and $L_2 = \emptyset$. The probabilities for the $L_2$ case can be obtained by using Proposition 2 parts a and c. Given $f_{G,T}$ for $f_H$, $f_S$ in Table 2, the appropriate value for $c$ in Theorem 1 is $c_1 = 1.29$ and $c_2 = 0.60$ respectively. As in Example 4, for a given $N$, which denotes the number of pairs in a desired code, let $\alpha_N = 1 - \frac{1}{2N}$ and let $g_N \equiv g(\alpha_N, N)$ be given by (5.6) when $T = 310°\text{K}$.

**Example 5.** *In Figure 3, using the left y-axis, the blue and yellow graphs plot the points $(t_1, \log_{10} N_1)$ and $(t_2, \log_{10} N_2)$ respectively that minimize even $t_i$ for given $N_i$ subject to the constraint that $Bad_{L_i}(N, t_i, g_N, c_i) \leqslant \frac{1}{2}$ as given in (6.15). The right y-axis gives the corresponding points $(t_i, g_{N_i})$. Thus, $t_i$ is the* theoretically computed sufficient length of DNA strands *such that there is a $DNA_{f_{G,T}}(t_i, L_i, g_i)$ code that self-assembles with fidelity $\alpha_i$ as defined in Definition 10. Note that $t_i$ is the sufficient length of DNA strands such that there is a $DNA_{f_{G,T}}(n_i, L_i, g_N)$ code that self-assembles with an expected failure rate of one WC duplexes failing to form per all possible $2N$ WC duplexes in the code (of strand multiplicity two.)* F3

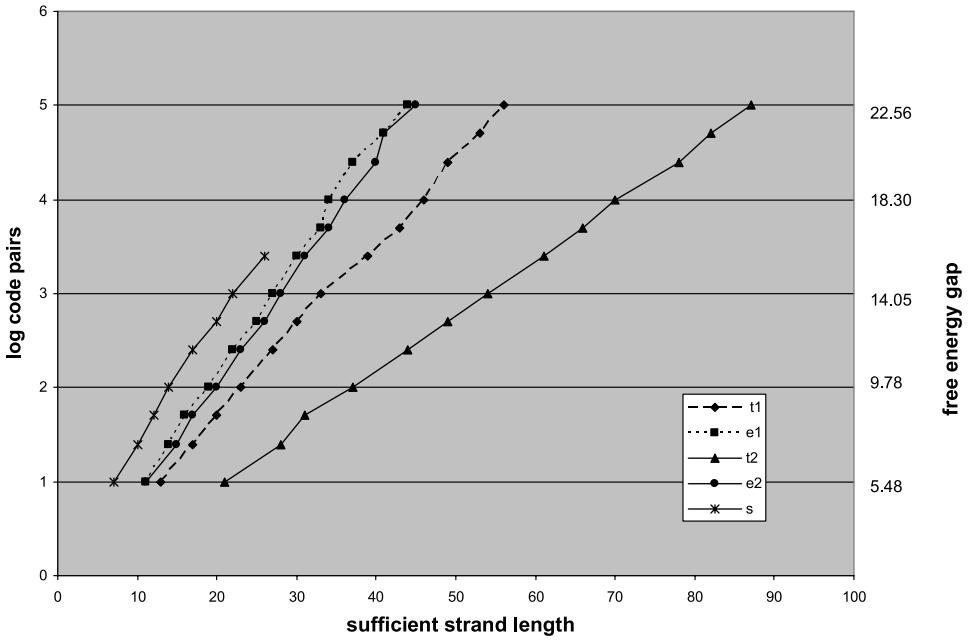


**FIG. 3.** Theoretical and empirical coding bound.

*To take a specific instance, if* $N_1 = 1000$*, then* $\alpha_N = 0.9995$*, so* $g_N = 14.05$*. Thus, a sufficient strand length is* $t_1 = 33$*, because*

$$Bad_{L_1}(1000, 33, 14.05, 1.29) = 0.42$$

*while*

$$Bad_{L_1}(1000, 32, 14.05, 1.29) = 1.02.$$

*However,* $t_2 = 54$*, because*

$$Bad_{L_2}(1000, 54, 14.05, 1.29) = 0.23$$

*while*

$$Bad_{L_2}(1000, 53, 14.05, 1.29) = 0.64.$$

*This difference depends on the difference between* $L_1$ *and* $L_2$ *and is discussed in the Conclusion section below.*

**Example 6.** *Given* $f_{G,T}$ *and* $L_i$*,* $\alpha_N$ *and* $g_N$ *as in Example 5, the pink and light blue graphs in Figure 3 respectively plot the points* $(e_1, \log_{10} N_1)$ *and* $(e_2, \log_{10} N_2)$ *that minimize strand length* $e_i$ *for given* $N_i$ *subject to the constraint that*

$$\beta_{i,g}(x, x) + (2N - 1)\beta_{i,g}(x, y) \leqslant \frac{1}{2}$$

*as given in (6.11) and where*

$$\beta_{i,g}(x, x) + (2N - 1)\beta_{i,g}(x, y)$$

*was empirically estimated by performing* $10^6$ *trials sampling from* $DNA(e_i, L_i)$*. Thus* $e_i$ *empirically estimates the sufficient length of DNA strands such that there is* $DNA_{f_{G,T}}(e_i, L_i, g_N)$ *code with N pairs that* self-assembles with fidelity $\alpha_N$ *as defined in Definition 10. To take a specific instance, if* $N_1 = 1000$*,*

*then the empirically estimated sufficient strand length is* $e_1 = 27$ *for strands that don't contain 2-strings in* $\{AA, TT, AT, TA\}$. *Note that for* $N_2 = 1000$, *the empirically estimated sufficient strand length is* $e_1 = 28$ *for strands with no restrictions on their 2-strings. In general, the graphs indicate that codes taken from* $DNA(n, L_1)$ *have slightly better emperically estimated random coding bounds than those taken from* $DNA(n, L_2)$.

**Example 7.** *Given* $f_{G,T}$ *and* $L_1$, $\alpha_N$ *and* $g_N$ *as in Example 5, the maroon graph in Figure 3 plots the points* $(s, \log_{10} N)$ *that minimize strand length s for given N subject to the constraint that the DNA code software SynDCode (Bishop et al., 2006) has produced a* $DNA_{f_{G,T}}(s, L_1, g_N)$ *code.*

## 8. CONCLUSION

A new type of DNA code, called a $DNA_{f_{G,T}}(n, L, g)$ *free energy gap code* has been defined and a statistical thermodynamic probabilistic model for $DNA_{f_{G,T}}(n, L, g)$ code self-assembly has been given. Theoretical and empirical random coding lower bounds for $DNA_{f_{G,T}}(n, L, g)$ codes were obtained and used to determine sufficient DNA code design parameters needed to achieve experimental goals. It has been noted that small changes in sequence composition, e.g., whether the consecutive pairs, AA, TT, AT or TA appear in any code sequence makes a difference in the potential fidelity of a DNA code, but perhaps not such as wide a difference in the theoretical sufficient strand length bounds between the cases when $L_1 = \{AA, TT, AT, TA\}$ and $L_2 = \varnothing$ as is shown in Example 5. By observing the empirical results in Examples 6 and 7, it is clear the theoretical bound for the case $L_2 = \varnothing$ exhibited in Figure 3 by the yellow graph is poor in comparison to that given for the case $L_1 = \{AA, TT, AT, TA\}$ exhibited in Figure 3 by the blue graph. This seems to be because that in the $L_2 = \varnothing$ case, the $c = 0.60$ for Thereom 1 is far further from the average value 1.42 of $f_{G,T}$ over the stack pairs not in $L_2 = \varnothing$ than is the $c = 1.29$ in the $L_1 = \{AA, TT, AT, TA\}$ case from average value 1.59 of $f_{G,T}$ over the stack pairs not in $L_1 = \{AA, TT, AT, TA\}$.

Earlier work on free energy gap collections of oligos is summarized in Tulpan et al. (2005). The results presented here suggest that the free energy gaps for collections that are given there may be too small. As discussed in Example 5, the fidelity $\alpha$ of a code with $N$ pairs should be greater than $1 - \frac{1}{2N}$ if that code is to self-assemble with an expected failure rate of less than one WC duplex. Thus, if $N \geq 64$, it follows from (5.6) that the free energy gap must be greater than 8.95. However, none of the 19 collections given in Table 2 of Tulpan et al. (2005) with $N \geq 64$ have a free energy gap greater that 8.95. Moreover, the *SynDCode* data exhibited in Figure 3, indicates that for strands of length 16, codes larger than those listed in Tulpan et al. (2005) may be found. Other sequence constraints for collections of oligos are considered in Tulpan et al. (2005). *SynDCode* (Bishop et al., 2006) also allows for consideration for many of these same constraints. For example, the code S8-2 given in Tulpan et al. (2005) is nearly a $DNA_{f_{G,T}}(16, L, 7.85)$ code where $L = \{GGG, CCC\}$. It is not exactly such a code for several reasons. The two most significant reasons are as follows:

1. S8-2 doesn't constrain the what is called the *word-word crosshybridization potential* (because an underlying in assumption (Tulpan et al., 2005) is that the strands $x$ are fixed to a surface.)
2. S8-2 uses *Pairfold* (Andronescu et al., 2003) to measure the free energy gap while $DNA_{f_{G,T}}(16, L, 7.85)$ uses $\gamma_{f_{G,T}}$.

   As was noted earlier, $\gamma_{f_{G,T}}$ is more restrictive that *Pairfold* (see Example 3). Some of the most important additional constraints for $S8-2$ are:

   For $x$ in $S8-2$ that:

   a. $CC$ is not contained at the start or end of $x$ and therefore $GG$ is not contained at the start or end of $\overline{x}$.
   b. $G$ is not contained in $x$ and therefore $C$ is not contained in $\overline{x}$.
   c. $15.45 \leq \|x, \bar{x}\|_{f_{G,T}} - IP \leq 16.42$ where we are assuming $IP = 1.96$.

*SynDCode* can generate $DNA_{f_{G,T}}(n, L, g)$ with each of these additional constrains. In particular, in Figure 4, an example of a $DNA_{f_{G,T}}(16, \{GGG, CCC\}, 7.85)$ code is provided that satisfies conditions a–c. **F4**

| Code Strands | | | |
|---|---|---|---|
| TCCTAAACCATTCTTA, $S_1$ | TAAGAATGGTTTAGGA, $C_1$ | CATCAAAAAACTCAAT, $S_{50}$ | ATTGAGTTTTTTGATG, $C_{50}$ |
| ACACATACTACTTCAA, $S_2$ | TTGAAGTAGTATGTGT, $C_2$ | AAAACACTTCTTTACT, $S_{51}$ | AGTAAAGAAGTGTTTT, $C_{51}$ |
| TTTCATTTCCAAAAAC, $S_3$ | GTTTTTGGAAATGAAA, $C_3$ | TTTTTACCTTAACCTC, $S_{52}$ | GAGGTTAAGGTAAAAA, $C_{52}$ |
| TAATAAACACACTCCA, $S_4$ | TGGAGTGTGTTTATTA, $C_4$ | CAATCTCTCTTCATAC, $S_{53}$ | GTATGAAGAGAGATTG, $C_{53}$ |
| CACTCTCACATTAAAA, $S_5$ | TTTTAATGTGAGAGTG, $C_5$ | ACACTAAAAAAACAAC, $S_{54}$ | GTTGTTTTTTTAGTGT, $C_{54}$ |
| ACATCTTCCTATACAA, $S_6$ | TTGTATAGGAAGATGT, $C_6$ | ATCATAACACAAACTT, $S_{55}$ | AAGTTTGTGTTATGAT, $C_{55}$ |
| ATCCATCAAACATAAA, $S_7$ | TTTATGTTTGATGGAT, $C_7$ | CACACATCATTTTTTC, $S_{56}$ | GAAAAAATGATGTGTG, $C_{56}$ |
| TTTCTCAATCCTACTA, $S_8$ | TAGTAGGATTGAGAAA, $C_8$ | CTATCCACCTAAAAAA, $S_{57}$ | TTTTTTAGGTGGATAG, $C_{57}$ |
| TTATACACCTCACTTT, $S_9$ | AAAGTGAGGTGTATAA, $C_9$ | TATATCCTCTCCAATC, $S_{58}$ | GATTGGAGAGGATATA, $C_{58}$ |
| AAAATTCATCAACCAA, $S_{10}$ | TTGGTTGATGAATTTT, $C_{10}$ | CTACATATCTAACCAC, $S_{59}$ | GTGGTTAGATATGTAG, $C_{59}$ |
| CAACCACTTAATCTTC, $S_{11}$ | GAAGATTAAGTGGTTG, $C_{11}$ | CTTACTACCAAACTTC, $S_{60}$ | GAAGTTTGGTAGTAAG, $C_{60}$ |
| ACCTTTTCTAACTCAT, $S_{12}$ | ATGAGTTAGAAAAGGT, $C_{12}$ | AAAAAATCCACATTTC, $S_{61}$ | GAAATGTGGATTTTTT, $C_{61}$ |
| ACTTATAAACCTACCA, $S_{13}$ | TGGTAGGTTTATAAGT, $C_{13}$ | AATTCTCCACTTTAAC, $S_{62}$ | GTTAAAGTGGAGAATT, $C_{62}$ |
| ACTCACCTCAATATAC, $S_{14}$ | GTATATTGAGGTGAGT, $C_{14}$ | AATTTCAACTATCACA, $S_{63}$ | TGTGATAGTTGAAATT, $C_{63}$ |
| TCCATATTCATACCTC, $S_{15}$ | GAGGTATGAATATGGA, $C_{15}$ | CAAAAAAAATCTCCTC, $S_{64}$ | GAGGAGATTTTTTTTG, $C_{64}$ |
| CACAATTTTTACAACT, $S_{16}$ | AGTTGTAAAAATTGTG, $C_{16}$ | TCACTTTTACCATTAC, $S_{65}$ | GTAATGGTAAAAGTGA, $C_{65}$ |
| ACTCTTTCTTTTCCTT, $S_{17}$ | AAGGAAAAGAAAGAGT, $C_{17}$ | TACTCCAACATTTTAC, $S_{66}$ | GTAAAATGTTGGAGTA, $C_{66}$ |
| TCTCATTAACATACAC, $S_{18}$ | GTGTATGTTAATGAGA, $C_{18}$ | AAAACACCACAAATAA, $S_{67}$ | TTATTTGTGGTGTTTT, $C_{67}$ |
| CATTCTTACTCTCTCT, $S_{19}$ | AGAGAGAGTAAGAATG, $C_{19}$ | TAACTTTCTATCTCCA, $S_{68}$ | TGGAGATAGAAAGTTA, $C_{68}$ |
| CTCTCTCAACTTAATC, $S_{20}$ | GATTAAGTTGAGAGAG, $C_{20}$ | CTCCTCTTCTCTATTA, $S_{69}$ | TAATAGAGAAGAGGAG, $C_{69}$ |
| AAACTAACATCAACAT, $S_{21}$ | ATGTTGATGTTAGTTT, $C_{21}$ | ACCTAACAATACAATC, $S_{70}$ | GATTGTATTGTTAGGT, $C_{70}$ |
| ATTCAAATCTTCCATC, $S_{22}$ | GATGGAAGATTTGAAT, $C_{22}$ | TCCTAAAATATCACAC, $S_{71}$ | GTGTGATATTTTAGGA, $C_{71}$ |
| CTCTAACTACTACCTT, $S_{23}$ | AAGGTAGTAGTTAGAG, $C_{23}$ | ACCAAAATACTCTTAC, $S_{72}$ | GTAAGAGTATTTTGGT, $C_{72}$ |
| CTTCATCATTTCTCTT, $S_{24}$ | AAGAGAAATGATGAAG, $C_{24}$ | ACAACTTCTCAATTTT, $S_{73}$ | AAAATTGAGAAGTTGT, $C_{73}$ |
| ATTAACTCCTCCTAAA, $S_{25}$ | TTTAGGAGGAGTTAAT, $C_{25}$ | AACATTACCTAATCAC, $S_{74}$ | GTGATTAGGTAATGTT, $C_{74}$ |
| CAACATACAATTACCA, $S_{26}$ | TGGTAATTGTATGTTG, $C_{26}$ | ATAACTCATCTTTCAC, $S_{75}$ | GTGAAAGATGAGTTAT, $C_{75}$ |
| CTCTTACAAAACATCT, $S_{27}$ | AGATGTTTTGTAAGAG, $C_{27}$ | TACACAACATATCCTA, $S_{76}$ | TAGGATATGTTGTGTA, $C_{76}$ |
| AACCTCTAAAATCCTA, $S_{28}$ | TAGGATTTTAGAGGTT, $C_{28}$ | CATCCACTTTTTTTTT, $S_{77}$ | AAAAAAAAAGTGGATG, $C_{77}$ |
| CAAAAACCAACTTTAA, $S_{29}$ | TTAAAGTTGGTTTTTG, $C_{29}$ | CTTCTACATCTCAAAA, $S_{78}$ | TTTTGAGATGTAGAAG, $C_{78}$ |
| TCTCTAATCACCAATA, $S_{30}$ | TATTGGTGATTAGAGA, $C_{30}$ | CTAAATACCATCCAAC, $S_{79}$ | GTTGGATGGTATTTAG, $C_{79}$ |
| ATATTACCACATCCTT, $S_{31}$ | AAGGATGTGGTAATAT, $C_{31}$ | ACTATCCAATTAACAC, $S_{80}$ | GTGTTAATTGGATAGT, $C_{80}$ |
| CATTTCACTCAAATTT, $S_{32}$ | AAATTTGAGTGAAATG, $C_{32}$ | TTTTTTATCTCACACA, $S_{81}$ | TGTGTGAGATAAAAAA, $C_{81}$ |
| CTTTACTTCCACAATA, $S_{33}$ | TATTGTGGAAGTAAAG, $C_{33}$ | CTACTACACTACACAT, $S_{82}$ | ATGTGTAGTGTAGTAG, $C_{82}$ |
| CACATATCCAAATCTC, $S_{34}$ | GAGATTTGGATATGTG, $C_{34}$ | ATTTTCACATTCTACA, $S_{83}$ | TGTAGAATGTGAAAAT, $C_{83}$ |
| CACCATAAATCATCAT, $S_{35}$ | ATGATGATTTATGGTG, $C_{35}$ | CAATATTCTTACACCA, $S_{84}$ | TGGTGTAAGAATATTG, $C_{84}$ |
| TTCACAAACCTTAATT, $S_{36}$ | AATTAAGGTTTGTGAA, $C_{36}$ | TCATACTTTATCCTCA, $S_{85}$ | TGAGGATAAAGTATGA, $C_{85}$ |
| ACCTCTTTATTCAAAC, $S_{37}$ | GTTTGAATAAAGAGGT, $C_{37}$ | TTTCCAATACATCAAT, $S_{86}$ | ATTGATGTATTGGAAA, $C_{86}$ |
| TCCTTACACAAATAAC, $S_{38}$ | GTTATTTGTGTAAGGA, $C_{38}$ | AAATCAAATACCTCTC, $S_{87}$ | GAGAGGTATTTGATTT, $C_{87}$ |
| ACAAATCACTTAAACA, $S_{39}$ | TGTTTAAGTGATTTGT, $C_{39}$ | TCCATTCCTTCTAATA, $S_{88}$ | TATTAGAAGGAATGGA, $C_{88}$ |
| CTCCTTCAATTTTCAA, $S_{40}$ | TTGAAAATTGAAGGAG, $C_{40}$ | TCCTTATTACTCCTAC, $S_{89}$ | GTAGGAGTAATAAGGA, $C_{89}$ |
| ATCAACACTATTCATC, $S_{41}$ | GATGAATAGTGTTGAT, $C_{41}$ | CTTCTCTTATTTCACA, $S_{90}$ | TGTGAAATAAGAGAAG, $C_{90}$ |
| TTCTTTCCTCATTTTT, $S_{42}$ | AAAAATGAGGAAAGAA, $C_{42}$ | TACTTCCTTTTTCTTC, $S_{91}$ | GAAGAAAAAGGAAGTA, $C_{91}$ |
| TATCACTCATACATCT, $S_{43}$ | AGATGTATGAGTGATA, $C_{43}$ | ACCACACACTATATAT, $S_{92}$ | ATATATAGTGTGTGGT, $C_{92}$ |
| CAACCATATCAAAAAC, $S_{44}$ | GTTTTTGATATGGTTG, $C_{44}$ | TACTTCACTAATTCCT, $S_{93}$ | AGGAATTAGTGAAGTA, $C_{93}$ |
| AATATCCTTTCTCACT, $S_{45}$ | AGTGAGAAAGGATATT, $C_{45}$ | CTTCAACAACTAACTA, $S_{94}$ | TAGTTAGTTGTTGAAG, $C_{94}$ |
| AAACTATTTCCACTCT, $S_{46}$ | AGAGTGGAAATAGTTT, $C_{46}$ | ACACCTATTATTCTCT, $S_{95}$ | AGAGAATAATAGGTGT, $C_{95}$ |
| TTTTTCATTCACCTTT, $S_{47}$ | AAAGGTGAATGAAAAA, $C_{47}$ | ACAATCAATTCTACTC, $S_{96}$ | GAGTAGAATTGATTGT, $C_{96}$ |
| ATAACCAACCTACATT, $S_{48}$ | AATGTAGGTTGGTTAT, $C_{48}$ | CTATCAAACTCCTTTT, $S_{97}$ | AAAAGGAGTTTGATAG, $C_{97}$ |
| ACCATTTTTATTCCAT, $S_{49}$ | ATGGAATAAAAATGGT, $C_{49}$ | | |

**FIG. 4.** $DNA_{f_{G,T}}(16, \{GGG, CCC\}, 7.85)$ code with additional constraints.

Thus considering that $DNA_{f_{G,T}}(n, L, g)$ codes do not ignore word-word crosshybrization potential and use a more restrictive measure free energy gap, the code given in Figure 4 is much more restrictive than $S8-2$. The *number of strands* in $S8-2$ is 80 while *the number of pairs of strands* in the exhibited $DNA_{f_{G,T}}(16, \{GGG, CCC\}, 7.85)$ is $N = 97$. It should be noted that there are other bonding specificity constraints considered in $S8-2$ that are not considered in the code given in Figure 4 and, in light of the random coding bound data $t_1$ and $e_1$ in Figure 3, $S8-2$ is a good design.

## ACKNOWLEDGMENTS

We would like to thank Dan Tulpan and the BETA Lab at the University of British Columbia for providing us with Pairfold software and many helpful discussions. Anthony J. Macula would like to acknowledge the support he receives from is family and most especially from his partner Jean T. Hennessey. This work has been supported by AFOSR FA8750-07-C-0089 and NSF-DMS/BIO UBM 0436298.

## REFERENCES

Andronescu, M., Aguirre-Hernandez, R., Condon, A., et al. 2003. RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.* 31, 3416–3422. Available at: *www.rnasoft.com*. Accessed August 1, 2007.

Bishop, M., Macula, A., and Renz, T. 2006. *SynDCode Suite*. Available at: *http://syndcode.geneseo.edu/*. Accessed August 1, 2007.

Braich, R., Chelyapov, N., Johnson, E., et al. 2002. Solution of a 20-variable 3-SAT problem on a DNA computer. *Sciencexpress* 296, 499–502.

Cai, H., White, P., Torney, D., et al. 2000. Flow cytometry–based minisequencing: a new platform for high throughput single nucleotide polymorphism scoring. *Genomics* 66, 135–143.

Chen, J., Deaton, R., and Garzon, M. 2006. Characterization of non-crosshybridizing DNA oligonucleotides manufactured *in vitro*. *Nat. Comput.* 5, 65–181.

Dirks, M., Lin, M., Winfree, E., et al. 2004. Paradigms for computational nucleic acid design. *Nucleic Acids Res.* 32, 1392–1403.

Dirks, R., Bois, J., Schaeffer, J., et al. 2007. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.* 49, 65–88.

D'yachkov, A., Erdös, P., Macula, A., et al. 2003. Exordium for DNA codes. *J. Combin. Optim.* 7, 369–380.

D'yachkov, A., Macula, A., Pogozelski, W., et al. 2005a. A weighted insertion deletion stacked pair thermodynamic metric for DNA codes. *Lect. Notes Comput. Sci.* 3384, 90–103.

Dyachkov, A., Macula, A., Pogozelski, W., et al. 2006. New t-gap insertion-deletion like metrics for DNA hybridization thermodynamic modeling. *J. Comput. Biol.* 13, 866–881.

D'yachkov, A., Vilenkin, P., Ismagilov, I., et al. 2005b. On DNA codes. *Probl. Inform. Transmiss.* 41, 349–367.

Eason, R., Pourmand, N., Tongprasit, W., et al. 2004. Characterization of synthetic DNA bar codes in *Saccharomyces cerevisiae* gene-deletion strains. *Proc. Natl. Acad. Sci. USA* 101, 11046–11051.

Fish, D., Horne, M., Searles, R., et al. 2007. Multiplex SNP discrimination. *Biophys. J.* 92, 89–92.

Hardenbol, P., Baner, J., Jain, M., et al. 2003. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* 6, 673–678.

Horne, M., Fish, D., and Benight, A. 2006. Statistical thermodynamics and kinetics of DNA multiplex hybridization reactions. *Biophys. J.* 91, 4133–4153.

Kaderali, L., Deshpande, A., Nolan, J., et al. 2003. Primer-design for multiplexed genotyping. *Nucleic Acids Res.* 31, 1796–1802.

Mathews, D., Zuker, M., and Turner, D. 2006. RNAstructure 4.2. Available at: *http://rna.chem.rochester.edu*. Accessed August 1, 2007.

Penchovsky, R., and Ackermann, J. 2003. DNA library design for molecular computation. *J. Comput. Biol.* 10, 215–229.

Rose, J., Deaton, R., Francescetti, D., et al. 1999. A statistical mechanical treatment of error in the annealing biostep of DNA computation. *Proc. Genet. Evol. Comput. Conf.* 2, 829–1834.

Rose, J., Deaton, R., and Suyama, A. 2004. Statistical thermodynamic analysis and design of DNA-based computers. *Nat. Comput.* 3, 443–459.

SantaLucia, J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95, 1460–1465.

Shortreed, M., Chang, S., and Hong, D. 2005. A thermodynamic approach to designing structure-free combinatorial DNA word sets. *Nucleic Acids Res.* 33, 4965–4977.

Tulpan, D., Andronescu, M., Chang, S., et al. 2005. Thermodynamically based DNA strand design. *Nucleic Acids Res.* 33, 4951–4964.

Valignat, M., Theodoly, O., Crocker, J., et al. 2005. Reversible self-assembly and directed assembly of DNA-linked micrometer-sized colloids. *Proc. Natl. Acad. Sci. USA* 102, 4225–4229.

Zhang, Y., Hammer, D., and Graves, D. 2005. Competitive hybridization kinetics reveals unexpected behavior patterns. *Biophys. J.* 89, 2950–2959.

Zuker, M., Mathews, D., and Turner, D. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski, J., and Clark, B.F.C., eds., *RNA Biochemistry and Biotechnology, NATO ASI Series*. Kluwer Academic Publishers, Amsterdam. Available at: *www.bioinfo.rpi.edu/∼zukerm*. Accessed August 1, 2007.

Address reprint requests to:
*Dr. Anthony J. Macula*
*Biomathematics Group*
*SUNY Geneseo*
*Geneseo, NY 14454*

*E-mail:* macula@geneseo.edu